10th ISCA Speech Synthesis Workshop
20-22 September 2019, Vienna, Austria

Statistical parametric synthesis of budgerigar songs

Lorenz Gutscher[1,2], *Michael Pucher*[1], *Carina Lozo*[1], *Marisa Hoeschele*[1,3], *Daniel C. Mann*[3,4]

[1]Acoustics Research Institute, Austrian Academy of Sciences
[2]SPSC, Graz University of Technology, Austria
[3]Department of Cognitive Biology, University of Vienna, Austria
[4]The Graduate Center of the City University of New York, USA

lorenz_gutscher@yahoo.de,
{michael.pucher,carina.lozo,marisa.hoeschele}@oeaw.ac.at,dmann@gradcenter.cuny.edu

Abstract

In this paper we present the synthesis of budgerigar songs with Hidden Markov Models (HMMs) and the HMM-based Speech Synthesis System (HTS). Budgerigars can produce complex and diverse sounds that are difficult to categorize. We adapted techniques that are commonly used in the area of speech synthesis so that we can use them for the synthesis of budgerigar songs. To segment the recordings, the songs are broken down into phrases, which are sounds separated by silence. Complex phrases furthermore can be subdivided into smaller units and then be clustered to identify recurring elements. These element categories along with additional contextual information are used together to enhance the training and synthesis. Overall, the aim of the process is to offer an interface that generates new sequences and compositions of bird songs based on user input, consisting of the desired song structure and contextual information. Finally, an objective evaluation comparing the synthesized output to the natural recording is performed, and a subjective evaluation with human listeners shows that they prefer resynthesized over natural recordings and that they perceive no significant differences in terms of naturalness between natural, resynthesized, and synthesized versions[1].

Index Terms: speech synthesis, bird song, bioacoustics, HMM-based synthesis

1. Introduction

Despite the progress in synthesis methods for sounds such as speech and musical instruments, little progress has been made in the synthesis of other sounds. One example of a sound type that lags behind in synthesis methods is that of animal sounds [2]. Studying the synthesis of animal sounds can help the exploration of complexity in animal communication systems and the search for precursors of music and speech [3]. Parrots and songbirds are - like human beings - vocal learners and need experience with other vocalizing members of their species to develop more complex vocalizations [4]. Being able to produce realistic sounds by synthesis provides an opportunity to design perceptual experiments with parrots and find out more about the features that they use and require to identify and discriminate among vocalizations. At the same time studying the synthesis of animal sounds can be useful for artistic purposes such as music and film productions, virtual reality, and game design.

To use Hidden Markov Models (HMMs) for the resynthesis of bird sounds is presented in [2], where the songs of chaffinches are segmented by the supervision of experts and manually labeled. In this paper we additionally present the implementation of an automatic segmentation algorithm, the categorization process to identify recurring elements, an objective evaluation of the synthesized samples, and the outcome of a subjective listening test. In comparison to chaffinches, budgerigars have a vast repertoire of different sounds that can even include imitations of human speech [5]. In particular, budgerigar songs are highly variable, and their phrases rarely reoccur. Normally they vocalize in groups, which makes the achievement of high-quality recordings that only include songs from one specimen very hard. Because of the limited database available and the possibility to use speaker adaptive training the HMM-based Speech Synthesis System (HTS) was chosen. The use of synthesis systems that are based on deep neural networks seem very promising (e.g., [6]) and are part of future work but will need much more training data beforehand. The automatic segmentation of acoustic events together with the clustering of elements present a novel approach for the full synthesis of budgerigar songs. To our knowledge, there was no work with budgerigar synthesis made so far partly due to their complex songs and the difficulty to obtain high-quality recordings.

[1]Parts of this paper have been published in a master thesis [1].

2. Sound production and recording

The underlying vocal apparatus in budgerigars is close to the mammalian one, with a few critical differences. Different models have been developed that describe the mechanics of bird sound production, however the following seems the most promising [7]: Air is being pressed out from air sacks through the bronchi and syrinx, where tissues (labia) are stimulated so that they vibrate. The sound then propagates to the trachea and the larynx. In contrast to mammals, where the larynx is the primary source of sound generation, trachea and larynx operate more like a variable filter, while the syrinx is the primary source [8]. The budgerigar named "Puck" whose recordings were used for this paper was recorded in the Budgerigar Laboratory of the Department of Cognitive Biology at the University of Vienna (see Figure 1). The files were recorded as 48 kHz WAVE files with 16 bits per sample. To decrease reverberation a shotgun microphone was used, placed as close to the animal as possible. Putting a microphone close to a bird may change its behavior, but with time, some birds like Puck habituated to the microphone and continued to sing.

3. Methodology

3.1. Segmentation

On the experimental basis of a segmentation script [9] the songs are cut into smaller units. Boundaries between elements are

10.21437/SSW.2019-23

Figure 1: *Budgerigar recorded with a shotgun microphone*

defined based on rapid changes of the parameters in the audio files (amplitude, fundamental frequency (F_0), Wiener Entropy). Here these segmentation rules are defined and applied to recordings of one budgerigar specimen. Detailed information of the actual segmentation algorithm can be found in [9], [10].

As bird species have very different vocalizations, attempts to standardize the units and their names now will be made to clarify it for the work with the current data set of budgerigar songs and the training of the toolkit. Songs are the biggest unit and can be divided into phrases, of which 7 different phrase types will be used here: contact call-like[2], long harmonic, short harmonic, alarm, noisy, click, unknown [11].

As phrases are isolated by short amounts of silence before and after them, a temporal division is sufficient to divide the songs into phrases and silent parts. Morphological methods (segmentation decisions based on specific parameter changes) then were used to segment contact call-like phrases into smaller units called elements [9], [12]. If contact call-like phrases have short parts of silence within a phrase, we additionally split up the phrase at this point and call them subphrases (see Figure 2).

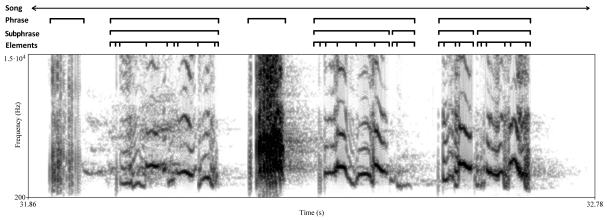

Figure 2: *An example of unit division of a budgerigar song*

Throughout all the recordings around 48 000 segment boundaries were automatically detected by the algorithm. To avoid very short elements the minimum duration for an element was augmented to 5 ms. Yet the number of existing phones of budgerigar sounds is not known, which is why the elements within a contact call-like do not indicate any repetitive occurrences in our data set, which will be the task that is being described in the following section.

3.2. Clustering of elements

We subdivided the elements into voiced and unvoiced categories by observing the frequency information at the center frame of each element. Using the F_0 extraction in Praat [13] the fundamental frequency of the elements was extracted. If the center frame contained a fundamental frequency it was labeled as a voiced element, if not it was labeled unvoiced[3]. Afterwards, we clustered the voiced and unvoiced elements into further subcategories using Gaussian Mixture Modelling for Model-Based Clustering [14], [15]. With the clustering a reduction of more than 30 000 elements to 11 groups of voiced (v1, v2, ..., v10, v11) and 9 groups of unvoiced (uA, uB, ..., uI) elements is achieved.

3.2.1. Voiced elements

The data vectors used for the voiced elements consist of the first 12 coefficients of the 34^{th} order mel-cepstral analysis[4] as well as the energy, all measured on the center frame of each element. The reduction to 12 coefficients neglects information about the fine structure of the samples and achieves a preferable lower dimension in the clustering process. Additionally, the logarithmic fundamental frequency and Wiener Entropy[5] are added to the data vector, so that each vector ends up with a dimension of [1 x 15]. Direct use of these parameters would result in a domination of high values. To solve this problem, the data needs to be scaled[6], so that all parameters have the same range of numbers and can be compared. In an additional attempt to get a higher distinction between element groups, the information about fundamental frequency is weighted higher to have a heavier judicial effect in the clustering process. The calculation result and plot of Bayesian Information Criteria (BIC) values for different number of element groups can be seen in Figure 3 respectively.

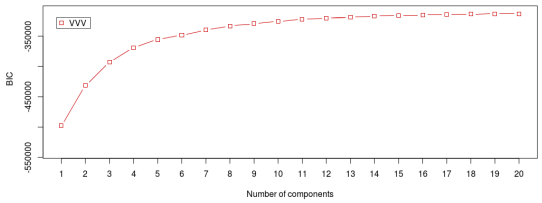

Figure 3: *BIC value for models with different numbers of element groups (*1-20*) for voiced elements (n =* 17 533*)*

The improvement of the BIC for models with more element groups is rather small. For an optimum solution we could see a rise to a maximum followed by a descent which is not the case for this data set. To choose the number of element groups, we tried to avoid both having too many groups containing rather similar elements and having only few element groups containing very distinct elements within the same element group. The chosen model is one with 11 element groups and an ellipsoidal distribution and varying volume, shape, and orientation (VVV).

3.2.2. Unvoiced elements

The unvoiced observation vectors have no F_0 information and therefore have a dimension of [1 x 14]. The result of the BIC estimation over different element group sizes can be seen in Figure 4 for the case of unvoiced elements. The chosen element group size follows the same procedure as above and is

[2]This phrase type will also contain compound phrases, that are combinations of different phrase types [9]

[3]Voicing threshold: 0.45, octave cost: 0.04, octave jump cost: 0.15, voiced/unvoiced cost: 0.04

[4]Parameter settings in HTS: FREQWARP = 0.55, GAMMA = 0

[5]This feature is not computed per frame but for each whole element

[6]This is done by subtracting the mean and division through standard deviation

one with 9 element groups and an ellipsoidal distribution and varying volume, shape, and orientation (VVV).

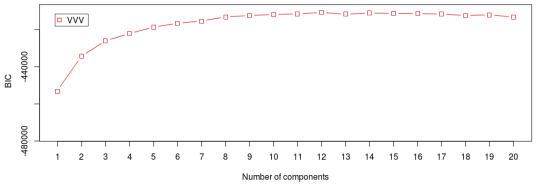

Figure 4: *BIC value for models with different numbers of element groups* (1-20) *for unvoiced elements* (n = 14 005)

3.3. HMM-based acoustic modeling

For modeling the bird songs, we started from a toolkit for opera singing synthesis [16] that was developed on the basis of a HMM-based singing synthesis system [17].

As a parametric representation of the spectral information mel-cepstral coefficients were used for training and testing. The Mel Log Spectrum Approximation (MLSA) vocoder was used to synthesize the sounds generated from statistical models and to resynthesize songs from mel-cepstral coefficients and F_0 parameters [18].

To cope with contextual variation, we use decision-tree based context clustering on the data set. This is done by defining a set of questions that matches with our available information of the songs. We had additional information about behavioral context, such as whether the bird performed a head bobbing display or to whom the song was directed. As it is not known which information will be useful in the final clustering, the aim was to provide as much contextual information as possible, which resulted in the following additional data for each element or phrase:

- previous/current/next element identity
- position of an element in current subphrase (forward/backward)
- previous/current/next element is voiced or not
- number of elements in previous/current/next subphrase
- position of the current subphrase in current phrase (forward/backward)
- number of subphrases in previous/current/next phrase
- number of elements in previous/current/next phrase
- undirected, male directed, inanimate directed, mixture of all three
- head movement in current song (no, yes, unknown)

Because the segmentation and clustering are not accurate for all phrases and elements, especially longer files produced computation errors during the training and had to be deleted from the corpus. After removing the error causing files, the training corpus consists of 62 WAVE files (each between 3 and 44 seconds long) and a total duration of 21 minutes and 2 seconds (out of 27 minutes 13 seconds). In a typical scenario the sounds to be synthesized as well as their order (e.g., "silence,v2,v4,uA,silence,v7") have to be specified in a label file. Most of the additional context information can be calculated automatically, while factors like behavioral descriptions need to be specified by the user. The toolkit will then compute the most likely acoustical models and outputs a WAVE file with the corresponding sound.

4. Analysis

Figure 5 shows the comparison of spectrograms between the natural recording and two synthesized versions. The synthesized version with natural duration offers good comparability to the natural recording, whereas the synthesized version with synthesized duration is a full synthesis, that uses the duration for each segment from the trained duration model. The fundamental frequency is emphasized on those parts where it is detected. In both examples the synthesized versions appear highly similar to the natural ones. The harmonics follow the contour of the ones obtained from the natural recording to some degree, but miss parts of the fine structure, which makes the synthesized versions sound a bit whistle-like and lack some noisiness. Duration modeling works well altogether, but as expected the aligned versions have a higher level of conformity to the original. It can be seen that harmonic sounds, which are labeled as unvoiced elements are synthesized with less energy and broadband (see the "uG" element in Figure 5).

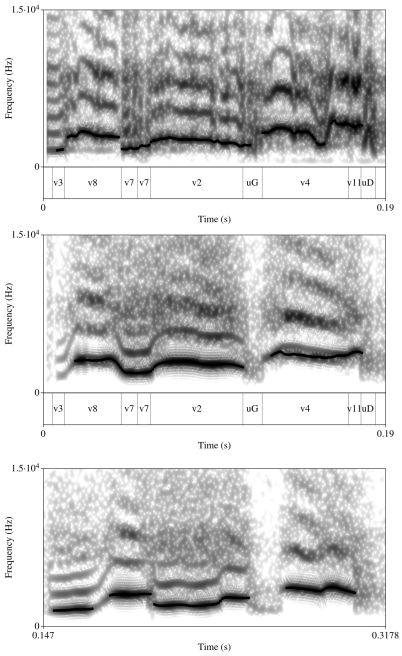

Figure 5: *Spectrogram of natural version (top), synthesized version with natural duration (middle) and synthesized version with synthesized duration (bottom)*

Figure 6 shows the estimation of the fundamental frequency of the natural version and of a synthesized version with natural duration. The fundamental frequency of the synthesis with natural duration (green line) follows the overall contour of the fundamental frequency of the natural version (black line) but misses some quick variations at the end of the song.

5. Evaluation

5.1. Objective evaluation

To compare the synthesized versions with the natural recordings mel-cepstral distortion is used where mel-cepstral coefficients are compared using dynamic time warping and a distance score. A high distance score indicates that the two data vectors are very different from each other, whereas a low score signifies high

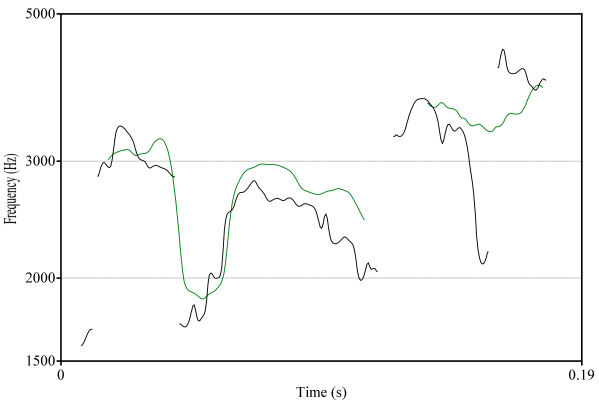

Figure 6: *F_0 comparison of natural version (black line) and synthesized version with natural duration (green line)*

conformity [19]. The distance scores between the methods are shown in Table 1. The first two columns use the duration from the natural version, the third column uses synthesized duration.

For the resynthesized version mel-cepstral coefficients and F_0 are extracted from the natural recordings and the MLSA vocoder is used to synthesize songs from these parameters. We can see minor differences of amplitude in the waveform (see Figure 7) that arise from the source-filter synthesis technique that is being used. Investigation of the spectrogram also reveals a decrease of overtones.

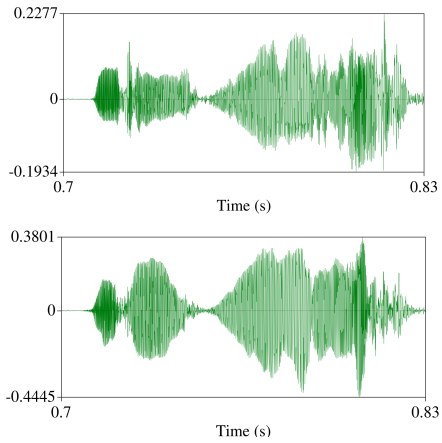

Figure 7: *Waveform of natural (top) and resynthesized (bottom) song*

Natural and resynthesized versions match best for all files, as no statistical modeling is involved. The synthesized version with synthesized duration has the highest distance score. Compared to the natural recordings, we can see that the synthesis with natural duration is in between the distance scores of the resynthesized and the fully synthesized version.

5.2. Subjective evaluation

In addition to the objective evaluation we also performed a subjective listening test[7]. Three different versions were used in the evaluation:

[7]The samples used in the listening test can be found on `https://speech.kfs.oeaw.ac.at/budgiessw10/`

Table 1: *Distance between natural (Nat.), resynthesized (Resyn.) and synthesized (Syn.) versions*

File	Nat. - Resyn. Nat. duration	Nat. - Syn. Nat. duration	Nat. - Syn. Syn. duration
10	0.78	1.24	1.48
15	0.81	1.20	1.78
17	0.84	1.45	1.68
99	0.81	1.53	1.64

1. *Natural* versions of budgerigar songs.
2. *Resynthesized* versions where mel-cepstral coefficient and F_0 are extracted from the natural recordings and the MLSA vocoder is used to synthesize songs from these parameters.
3. *Synthesized* versions where all parameters (mel-cepstral coefficients, F_0, duration) are predicted from HMMs given a sequence of input labels, and then synthesized with the MLSA vocoder.

The three different methods were evaluated by 22 listeners (8 ♀, 12 ♂, 2 NA). Subjective evaluation by humans is relevant for applications in computer games or virtual reality.

A survey was set up with [20] and participants were recruited via an email inviting them to take part in the study. Prior to completing the task, the participant was presented with an excerpt of a natural budgerigar song. The participants were employees of the Acoustics Research Institute Vienna, all familiar with listening tests regarding synthesis, but only a few had special knowledge about bird vocalization. Participants then rated natural and synthesized versions on their naturalness. Each trial consisted of a screen with 2 versions of one of 7 songs that the participant could play back as many times as they liked (all 3 combinations of natural, resynthesized, synthesized). With all possible combinations of the versions there were $((3 * 2)/2) * 7 = 21$ comparisons for each listener. Participants were then asked to rate each song in regard to their naturalness by moving a sliding bar from 0% to 100% that was labeled "künstlich (artificial)" and "natürlich (natural)" on its ends. A pairwise comparison of the two samples followed, where the listener was forced to choose the sample they liked better. This way we can pull together the results of the human listening test and a place preference test [21] with budgerigars that is currently in work.

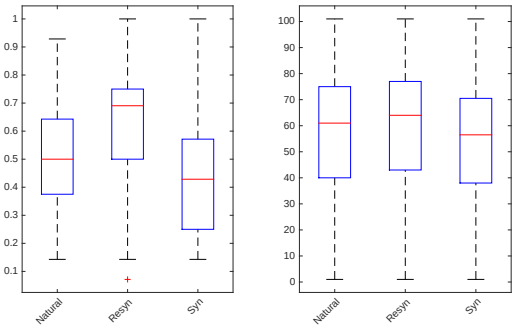

Figure 8: *Results of pairwise comparison (left) and naturalness rating (right).*

Figure 8 (left) shows the pairwise comparison score between the two methods for each listener. As can be seen from

Figure 8 (left) the resynthesized versions are judged slightly better than the natural samples and the synthesized ones, with differences being statistically significant between natural and resynthesized samples ($p < 0.004$) and between resynthesized and synthesized samples ($p < 0.001$) according to a Wilcoxon rank sum test. Interestingly there were no significant differences between natural and (fully) synthesized versions.

Concerning the ratings of naturalness, no significant differences can be found between natural, resynthesized, and synthesized samples.

6. Discussion

A synthesis toolkit based on Hidden Markov Models (HMMs) was developed, that produces budgerigar vocalizations from a user input file. The toolkit gives the possibility to conduct further experiments with budgerigars to learn more about their preferences and the importance of different syntax/patterns. Further research with budgerigar vocal production and their perception of their own vocalizations could help clarify how many element groups should be used to accurately synthesize budgerigar vocalizations. In addition, using the vocalizations of more individual budgerigars could increase the generalizability to the budgerigar species as a whole. Finally, the HMM-based Speech Synthesis System offers many parameters that can be adapted and experimented with to increase the naturalness of the result. The incorporation of vibrato and tremolo features is very successful in retaining spectral details and rapid volume changes [2] and might increase the vividness of areas where tremolo and vibrato appear.

Behavioral experiments with budgerigars could evaluate, whether the representation of the resynthesized samples actually seems natural to the birds. This test is currently in progress by using a place preference test (e.g., [21]) in a setup already familiar to the birds of the Viennese budgie lab, where the recordings were made. The preference test allows the birds to choose between three different wooden perches that are each placed in front of a different speaker. By sitting on the perch in front of a given speaker, a sound will begin playing from that speaker. Two of the speakers are used for playback, while one always remains silent. The elapsed time that a bird sits on each perch is then measured and evaluated. Greater time spent on a perch is thought to reflect greater preference for that sound. In this way, it is possible to evaluate which methods of synthesis lead to greater preference in the birds.

7. Conclusion

We have shown how to synthesize budgerigar songs from symbolic input label sequences by Hidden Markov Models that are trained on a corpus of labeled songs. The song labeling, and clustering of elements can be done in a semi-automatic fashion.

The subjective evaluation showed that human listeners prefer resynthesized versions over natural and synthesized ones and that there are no significant differences between the perception of synthetic and natural songs for human listeners. The perception of human listeners is relevant for usage of such synthesizers in computer games or virtual reality. Currently the developed synthesizers are used in bioacoustics for investigating the structure of budgerigar songs. In these experiments, different scales in addition to the mel-scale will be investigated.

8. References

[1] L. Gutscher, "Recording, analysis, statistical modeling and synthesis of bird songs," Master's thesis, Graz University of Technology, 2019.

[2] J. Bonada, R. Lachlan, and M. Blaauw, "Bird song synthesis based on hidden markov models," in *Interspeech 2016*. ISCA, 2016.

[3] P. Marler, *Origins of music and speech: insights from animals*, S. B. Nils L. Wallin, Bjrn Merker, Ed. A Bradford Book, 2001.

[4] W. H. Thorpe, "The leaning of song patterns by birds, with especial reference to the song chaffinch fringilla coelebs," *Ibis*, vol. 100, pp. 535–570, 1958.

[5] M. L. Dent, E. F. Brittan-Powell, R. J. Dooling, and A. Pierce, "Perception of synthetic /ba/–/wa/ speech continuum by budgerigars (melopsittacus undulatus)," *The Journal of the Acoustical Society of America*, vol. 102, no. 3, pp. 1891–1897, 1997.

[6] M. Nishimura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Singing voice synthesis based on deep neural networks," in *INTERSPEECH*, 2016.

[7] G. Mindlin and R. Laje, *The Physics of Birdsong*. Springer Berlin Heidelberg, 2006.

[8] E. Bezzel and R. Prinzinger, *Ornithologie*. Stuttgart, Ulmer Verlag, 1990, UTB groe Reihe Nr.8051, 1990.

[9] D. C. Mann, "Stabilizing forces in acoustic cultural evolution: Comparing humans and birds," Ph.D. dissertation, The City University of New York, 2019.

[10] D. C. Mann, W. Fitch, H.-W. Tu, and M. Hoeschele, "The building squawks of life: Human-like segments in budgerigar warble, in prep." 2019.

[11] H.-W. Tu, E. W. Smith, and R. J. Dooling, "Acoustic and perceptual categories of vocal elements in the warble song of budgerigars (melopsittacus undulatus)." *Journal of Comparative Psychology*, vol. 125, no. 4, pp. 420–430, 2011.

[12] N. Thompson, K. LeDoux, and K. Moody, "A system for describing bird song units," *Bioacoustics*, vol. 5, no. 4, pp. 267–279, 1 1994.

[13] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.1.13)," 2009. [Online]. Available: http://www.praat.org

[14] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014, accessed: 02.12.2018. [Online]. Available: https://www.R-project.org

[15] L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery, "mclust 5: clustering, classification and density estimation using Gaussian finite mixture models," *The R Journal*, vol. 8, no. 1, pp. 205–233, 2016, accessed: 25.11.2018. [Online]. Available: https://journal.r-project.org/archive/2016-1/scrucca-fop-murphy-etal.pdf

[16] M. Pucher, F. Villavicencio, and J. Yamagishi, "Development and evaluation of a statistical parametric synthesis system for operatic singing in German," in *Speech Synthesis Workshop (SSW9)*, Sunnyvale, CA, 2016, pp. 64–69.

[17] Sinsy, "HMM-based singing voice synthesis system," http://sinsy.sourceforge.net/, 2013.

[18] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (MLSA) filter for speech synthesis," *Electronics and Communications in Japan (Part I: Communications)*, vol. 66, pp. 10 – 18, 02 1983.

[19] SPTK, "Speech signal processing toolkit (sptk)," 2015, accessed: 16.11.2018. [Online]. Available: http://sp-tk.sourceforge.net/

[20] D. Leiner, "SoSci Survey. Version 3.1.06−i," http://www.soscisurvey.de/, 2019.

[21] M. Hoeschele and D. L. Bowling, "Sex differences in rhythmic preferences in the budgerigar (melopsittacus undulatus): A comparative study with humans," *Frontiers in Psychology*, vol. 7, p. 1543, 2016.